Filtered-variate Prior Distributions for Local Smoothing of Histograms[1]

by

James M. Dickey[2], University of Minnesota
Jyh-Ming Jiang, University of Lowell

School of Statistics, University of Minnesota
Technical Reprt No. 526
January, 1989

[1] The authors are grateful for discussions with David Lane, and for early discussions with participants at the NBER-NSF Seminar on Bayesian Inference in Econometrics, Rutgers University, October 26-27, 1984, and the Fifth Workshop on Maximum Entropy Methods, University of Wyoming, August 5-8, 1985.

[2] Research supported in part by NSF Research Grant DMS-8614793.

Filtered-Variate Prior Distributions for Local Smoothing of Histograms

by
James M. Dickey and Jyh-Ming Jiang

Abstract. Bayesian inference for histogram data has suffered from a lack or scarcity of convenient prior distributions permitting the expression of prior knowledge that the underlying population frequencies vary locally smoothly over categories. A general theory of "filtered-variate" prior distributions is here developed in which the prior-random probability vector is an adjustable linear transform of a standard random probability vector, or equivalently, a weighted average of specified nonrandom probability vectors, with random weights. Bayes' theorem and posterior analyses are conveniently carried out in terms of the prior and posterior densities of the weights vector. A promising method of expert prior assessment is given for filtered-variate priors, based on a theory of typical outcome vectors and a general representation theory for filters on random probability vectors having mean-structured variance, that is, on random probability vectors for which the variance matrix is a suitable quadratic function of the mean. Our prior-assessment scheme allows a subject-matter expert merely to provide a list of several typical probability vectors, typical in their smoothness. The empirical moments of the list can then be emulated in the mean vector and variance matrix of the resulting prior distribution. The theory and methods are detailed for the family of filtered-variate Dirichlet prior distributions, for which the posterior distribution of the weights vector is a case of Dickey's (1983, *J. Amer. Statist. Assoc.*, 628-637) generalized Dirichlet distribution. (A filtered-variate Dirichlet is not a "mixture of Dirichlet distributions," since only a single Dirichlet distribution is involved.)

1. Introduction.

How can a statistician effectively model an expert's uncertainty to express the expert's prior belief in local smoothness for the unknown sampling probabilities of histogram categories? How can the joint prior distribution give high prior probability to the event that the sampling probabilities are "smooth," that adjacent categories have probabilities close in value? In other words, how can one arrange high prior correlations between neighboring category probabilities, or low prior variances on their differences? For example, the categories may be grouping intervals and their probabilities may be the integrals of smooth, but otherwise unknown, density function. This problem is important for a wide range of applications, from uses of one-way histogram data (Dickey 1968b, 1969), to medical diagnosis, optical image processing, and other uses of multidimensional histograms (Dickey 1968a, Vardi, Shepp, and Kaufman 1985).

The problem has been an important one for decades, at least. In its extreme form, with an infinite or continuous list of categories, it is known as the problem of Bayesian nonparametric inference, and it has been a major embarrassment for Bayesian statisticians (Savage 1970). A review of the literature would be overly lengthy here, but the reader may find interest in the discussions of issues in Dickey (1968a), Leonard (1978), Lenk (1988), and Diaconis and Freedman (1986).

In its finite form, the problem can be described as follows. A vector will be called a **probability vector** if each coordinate is nonnegative and the coordinates sum to unity. Denote the simplex of probability vectors in R^k by Δ^{k-1}. A prior distribution for unknown $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^T$ is needed that will have the following four properties, in addition to giving unit probability to Δ^{k-1}.

i. The family of prior distributions of $\boldsymbol{\theta}$ must be large enough to allow accurate expression of real predata personal uncertainties concerning $\boldsymbol{\theta}$.

ii. Local smoothness, in particular, must be available as a property in the family -- "available" to the extent that the amount of local smoothness can be meaningfully specified and interpreted in terms of prior experience and expert belief.

iii. Following the receipt of statistical data, Bayes' Theorem calculations can be carried out simply.

iv. The resulting posterior distribution of $\boldsymbol{\theta}$ must be tractable, in that one can easily compute interesting and inferentially useful summaries of the posterior distribution.

We will propose a family of prior distributions on Δ^{k-1} satisfying these requirements, together with a potentially practical method of assessment. To appreciate the difficulties and establish notation, consider i.i.d. sampling from a finite distribution having a mass function with probabilities $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^T \in \Delta^{k-1}$, say with probability θ_i assigned to category i, $P(x=i \mid \boldsymbol{\theta}) = \theta_i$, for $i = 1, \ldots, k$. Under arbitrary noninformative stopping (Raiffa and Schlaifer 1961, sec. 2.3), the category counts $\mathbf{n} = (n_1, \ldots, n_k)^T$, suffice for the likelihood function of a sample sequence $\mathbf{x} = (x_1, \ldots, x_N)^T$, $n_+ = N$, where $n_+ \equiv n_1 + \ldots + n_k$,

$$p(N, \mathbf{x} \mid \boldsymbol{\theta}) \propto \prod_{i=1}^{k} \theta_i^{n_i}, \tag{1.1}$$

where the proportionality is taken with respect to $\boldsymbol{\theta}$. For example, if the sample size N is prespecified, then $(\mathbf{n} \mid \boldsymbol{\theta}) \sim$ multinomial$(N, \boldsymbol{\theta})$. The usual conjugate family of prior distributions for likelihoods (1.1) is the **Dirichlet**, $\boldsymbol{\theta} \sim D(\mathbf{b})$, $\mathbf{b} = (b_1, \ldots, b_k)^T$ with each $0 \le b_i \le \infty$, having the density (if each $b_i > 0$),

$$p(\boldsymbol{\theta}) = B(\mathbf{b})^{-1} \prod_{i=1}^{k} \theta_i^{b_i - 1}, \tag{1.2}$$

for $\boldsymbol{\theta} \in \Delta^{k-1}$, where $B(\mathbf{b}) = [\prod \Gamma(b_i)] / \Gamma(b_+)$. The k coordinates here are constrained to sum to unity, $\theta_+ = 1$, and this is the same density function with respect to **any** $k-1$ coordinates, for example, $\theta_1, \ldots, \theta_{k-1}$. The resulting posterior distribution would again be Dirichlet, with the updated parameter vector $\mathbf{b} + \mathbf{n}$,

$$\boldsymbol{\theta} \mid \mathbf{n} \sim D(\mathbf{b} + \mathbf{n}). \tag{1.3}$$

The nonsmooth character of Dirichlet distributions is revealed by their moments. For $\mathbf{d} = (d_1, \ldots, d_k)^T$, the general $\mathbf{d}$th mixed moment of $\boldsymbol{\theta} \sim D(\mathbf{b})$ is

$$E \prod_{i=1}^{k} \theta_i^{d_i - 1} = h(\mathbf{d}; \mathbf{b}), \tag{1.4}$$

where $h(\mathbf{d}; \mathbf{b}) = B(\mathbf{b} + \mathbf{d}) / B(\mathbf{b})$. So the mean vector and variance matrix are, respectively,

$$E\,\boldsymbol{\theta} = \mathbf{w}, \tag{1.5}$$

where $\mathbf{w} = \mathbf{b} / b_+$, and

$$\mathrm{Var}(\boldsymbol{\theta}) = (b_+ + 1)^{-1} [\mathbf{Diag}(\mathbf{w}) - \mathbf{w}^T\mathbf{w}]. \tag{1.6}$$

Distributions on the probability simplex in which the first two moments are related proportionally through such a matrix quadratic function will be defined later to have a "mean-structured variance". We will obtain interesting properties of such distributions in general, but for now, merely note that every prior correlation from such a variance matrix is necessarily negative or zero,

$$\mathrm{Corr}(\theta_i, \theta_j) = -\left[\frac{w_i}{1-w_i} \cdot \frac{w_j}{1-w_j}\right]^{1/2}. \tag{1.7}$$

Since the posterior distribution (1.3) from the prior (1.2) is again Dirichlet, the posterior moments are similar in character to the prior moments. So, for one thing, the posterior correlation between every two category probabilities is again negative or zero. The Dirichlet posterior mean can be written as an estimate that is computed by shrinking the usual unbiased, maximum-likelihood estimate, $\hat{\boldsymbol{\theta}} = \mathbf{n} / n_+$, toward the prior mean point $\mathbf{w}$,

$$E(\boldsymbol{\theta} \mid \mathbf{n}) = (1-u)\,\hat{\boldsymbol{\theta}} + u\,\mathbf{w} \ , \qquad (1.8)$$

where $u = b_+ / (b_+ + n_+)$. If $\mathbf{w}$ is smoother than $\hat{\boldsymbol{\theta}}$, then so is the posterior mean, and thus we would have a "smoothed" estimate. However, as recognized by Good and Gaskins (1971, 1980), this smoothing is global, rather than local, in the sense that the adjacent probability differences are diminished to no greater extent than are the nonadjacent differences. Indeed, the effect of such global smoothing on a posterior mean difference, $E(\theta_i \mid \mathbf{n}) - E(\theta_j \mid \mathbf{n})$, depends only on the corresponding prior mean difference $w_i - w_j$ and not on the distance between categories $|i - j|$,

$$E(\theta_i \mid \mathbf{n}) - E(\theta_j \mid \mathbf{n}) = (1-u)(\hat{\theta}_i - \hat{\theta}_j) + u\,(w_i - w_j) \ . \qquad (1.9)$$

If the prior mean is smooth, $|w_i - w_j|$ in (1.9) is small for short distances $|i - j|$. But $|w_i - w_j|$ is also small, even zero, for various long distances. The weights $1-u$ and u are independent of i and j.

A similar story holds for the posterior mode considered as an estimate. The Dirichlet distribution $\boldsymbol{\theta} \sim D(\mathbf{b})$ (1.2) has the mode $\mathbf{w}^-$, if $b_+ > k$, where

$$\mathbf{w}^- = (\mathbf{b} - \mathbf{1}_k) / (b_+ - k) \ . \qquad (1.10)$$

We denote a vector of k unit coordinates by $\mathbf{1}_k = (1, \ldots, 1)^T$. Since the posterior distribution (1.3) is Dirichlet, it will have the mode, obtained from (1.10) by replacing $\mathbf{b}$ by $\mathbf{b} + \mathbf{n}$ and isolating terms,

$$\text{Mode}(\boldsymbol{\theta} \mid \mathbf{n}) = (1-u^-)\,\hat{\boldsymbol{\theta}} + u^-\,\mathbf{w}^- \ , \qquad (1.11)$$

where $u^- = (b_+ - k) / (b_+ - k + n_+)$. The similarity of this expression for the posterior mode to the expression for the posterior mean (1.8) shows that the mode, too, is merely a global smoothing. The mean and mode will be approximately the same for large n_+, since they are related by.

$$E(\boldsymbol{\theta}\,|\,\mathbf{n}) = [(b_+ + n_+ - k)/(b_+ + n_+)]\,\mathrm{Mode}(\boldsymbol{\theta}\,|\,\mathbf{n}) + [k/(b_+ + n_+)]\,(\mathbf{1}_k/k)\,. \quad (1.12)$$

By (1.12), the mean is (globally) smoother than the mode.

In Bayesian **normal** linear sampling with a conjugate normal prior, the posterior mean and mode are identical and, again, this is a convex linear combination of the prior mean and the usual maximum likelihood estimate. But the normal family of distributions is closed under linear operations on the random vector, and so such a prior distribution can be assigned an arbitrary prior covariance structure. In obvious notation,

$$E(\boldsymbol{\mu}\,|\,\hat{\boldsymbol{\mu}}) = (\mathbf{I}-\mathbf{U})\,\hat{\boldsymbol{\mu}} + \mathbf{U}\,E(\boldsymbol{\mu})\,, \quad (1.13)$$

with $\mathbf{I}-\mathbf{U} = \mathbf{V}_{\boldsymbol{\mu}}(\mathbf{V}_{\boldsymbol{\mu}}+\mathbf{V}_{\hat{\boldsymbol{\mu}}|\boldsymbol{\mu}})^{-1}$, $\mathbf{U} = \mathbf{V}_{\hat{\boldsymbol{\mu}}|\boldsymbol{\mu}}(\mathbf{V}_{\boldsymbol{\mu}}+\mathbf{V}_{\hat{\boldsymbol{\mu}}|\boldsymbol{\mu}})^{-1}$. As pointed out by Titterington (1986), the difference of coordinates can be written,

$$E(\mu_i\,|\,\hat{\boldsymbol{\mu}}) - E(\mu_j\,|\,\hat{\boldsymbol{\mu}}) = (\mathbf{D}_{i,j}\,\mathbf{V}_{\boldsymbol{\mu}})(\mathbf{V}_{\boldsymbol{\mu}}+\mathbf{V}_{\hat{\boldsymbol{\mu}}|\boldsymbol{\mu}})^{-1}\hat{\boldsymbol{\mu}}$$

$$+ (\mathbf{D}_{i,j}\,\mathbf{V}_{\hat{\boldsymbol{\mu}}|\boldsymbol{\mu}})(\mathbf{V}_{\boldsymbol{\mu}}+\mathbf{V}_{\hat{\boldsymbol{\mu}}|\boldsymbol{\mu}})^{-1}E(\boldsymbol{\mu})\,, \quad (1.14)$$

where the operator $\mathbf{D}_{i,j}$ produces the difference between the ith and jth row vectors, $\mathbf{D}_{i,j}\,\mathbf{V} = (v_{i1}-v_{j1}, \ldots, v_{ik}-v_{jk})$. By (1.14), we see that the normal-theory smoothing can be truly local, in that a gently varying prior variance-covariance would imply a small difference between the ith and jth row vectors of $\mathbf{V}_{\boldsymbol{\mu}}$ for a short distance $|i-j|$, and thereby, a small effect of $\hat{\boldsymbol{\mu}}$ in the first term of (1.14).

Leonard (1973), Lenk(1988), and others achieved local smoothing of histogram data by exploiting and modifying the conjugate-prior normal theory. A nonlinear (logistic) change of variable in a multivariate normal distribution was used to guarantee prior certainty for the event that all the category probabilities are nonnegative and sum to unity. For related methods, see Good and Gaskins (1971, 1980). As an alternative theory, we work directly with a linear transform, or "filter", of a Dirichlet or other mean-structured vector, the support set of which will lie, naturally, within the probability simplex. Any class of mean-structured distributions, however, is not closed under linear transformations of the vector variate. The density of a linear transform is complicated, and if such a density is used as the prior density for multinomial sampling, the

posterior density will be even more complicated. However, we shall succeed in using the distribution of a linear transform of a mean-structured vector as the prior distribution of the category probabilities by maintaining the underlying untransformed vector as the variable of integration in the prior and posterior density functions. Both the prior and the resulting posterior distributions will then be tractable. The posterior mean and the posterior mode will be computable as estimates, and other inferential summaries and properties of the prior and posterior distributions will be given.

We formalize, in Section 2, the concept of filtering a random vector to obtain a "filtered-variate" version of its distribution, which we can propose as an expression of prior local smoothness. In Section 3., we develop the filtered-variate Dirichlet family of prior distributions and their consequent posterior distributions. The corresponding posterior family will be a filtered-variate form of Dickey's (1983) generalized Dirichlet distributions. We also give, in Section 3, the inferences from the posterior distribution for smoothed estimation and for hypothesis-comparison.

A problem of concern in this research is how to choose a filtered-variate prior distribution. A promising line of thinking for the development of practical methods was discovered by viewing the filtered-variate Dirichlet distributions as a subclass of filtered-variate forms of distributions having mean-structured variance. The crucial property of such a distribution is that its low-order moments can be represented in terms of a limiting finite distribution on the set of column vectors of the filter matrix. This enables one to elicit expert prior opinion in the form of a list of typically smooth category-probability vectors, which can then be scale-transformed to serve as the column vectors of the filtering matrix, thereby, yielding a filtered-variate prior distribution having the same low order moments as the empirical moments of the typical list. Modifications of the method allow symmetric or other direct and model-driven choices regarding the prior moments. The general theory of distributions with mean-structured variance and their filtered-variate forms is developed in Section 4., and used in prior assessment in Section 5. We conclude in Section 6. with comments on the continuous-dimensional Dirichlet process as a prior and relations of our work to more general finite-mixture distribution problems.

2. Filtered-Variate Distributions.

We obtain a random probability vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^T$ by defining $\boldsymbol{\theta}$ as a linear transform of an assumed random probability vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)^T$. Let

$$\boldsymbol{\theta} = \mathbf{G}\boldsymbol{\alpha} \ , \tag{2.1}$$

where $\mathbf{G}$ (k×m) is a constant matrix and the vector $\boldsymbol{\alpha}$ has a specified distribution. Then $\boldsymbol{\theta}$ will be said to have a **filtered-variate** form of the distribution of $\boldsymbol{\alpha}$. For example, $\boldsymbol{\alpha}$ may be Dirichlet distributed,

$$\boldsymbol{\alpha} \sim D(\mathbf{a}) \ , \tag{2.2}$$

$\mathbf{a} = (a_1, \ldots, a_m)$, in which case, $\boldsymbol{\theta}$ has a **filtered-variate Dirichlet distribution**. (The distribution is, in no way, a "mixture of Dirichlet distributions," since there is only one Dirichlet distribution involved.) We will denote the distribution (2.1), (2.2) by

$$\boldsymbol{\theta} \sim F_{\mathbf{G}}D(\mathbf{a}) \ . \tag{2.3}$$

Essentially, the one-dimensional filtered-variate Dirichlet distribution (k=2) was studied by Bloch and Watson (1967), Dickey (1983), Jiang (1984, 1988), and Cifarelli and Regazzini (1988). The usual Dirichlet is the special case, $D(\mathbf{a}) \sim F_{\mathbf{I}}D(\mathbf{a})$.

It is instructive to interpret a filtered-variate distribution (2.1) in two alternate ways:

> I. Each coordinate of $\boldsymbol{\theta}$ is a linear combination of the coordinates of the random vector $\boldsymbol{\alpha}$,
>
> $$\theta_i = g_{i,1}\alpha_1 + \ldots + g_{i,m}\alpha_m \ , \tag{2.4}$$
>
> for i = 1, . . ., k.

This is a linear filter on α. Hence, we use the compound adjective "filtered-variate". The variate is filtered, rather than the distribution, as would be the case for a density transformed by an integral operator.

II. The vector θ is a weighted average of the fixed column vectors of G with random weights α. For the array of column vectors,

$$G = (g_1, \ldots, g_m), \tag{2.5}$$

write

$$\theta = \alpha_1 g_1 + \ldots + \alpha_m g_m . \tag{2.6}$$

Since θ will need to be a probability vector for every realization of α, and in particular for each unit coordinate point, $\alpha_j = 1$, $\alpha_{j'} = 0$ for all $j' \neq j$, we obtain the following property.

Lemma 2.1 In case the support of α includes the extreme points of Δ^{m-1}, the requirement that the support of θ (2.3) be contained in Δ^{k-1} implies that each column vector g_j of G must be a probability vector.

Thus, we will assume that all entries of G are nonnegative and each column of G sums to unity. That is, the fixed matrix G is what is called a (singly) stochastic matrix.

If the support set of the random weights α is the full probability simplex Δ^{m-1}, then the support set of random θ is the convex hull of the set of fixed column vectors of G, $H(G)$, a convex polytope and subset of Δ^{k-1}. The vertices or extreme points of $H(G)$ are column vectors of G, but not necessarily all the columns of G will be vertices of $H(G)$. The polytope $H(G)$ would be a complicated range to work with if one developed and used a density for θ. For example, in the relatively simple case of $k=m$ and nonsingular G, with $S = G^{-1}$, the density of $\theta_1, \ldots, \theta_{k-1}$, where $\theta \sim F_G D(a)$, is

$$p(\theta) = \left| S_{11} - S_{12} 1_{k-1}^T \right| \; B(a)^{-1} \prod_i \left(\sum_j s_{ij} \theta_j \right)^{a_i - 1}, \tag{2.7}$$

for $\boldsymbol{\theta}$ restricted to $H\!I(\mathbf{G}) = \{\boldsymbol{\theta} : \Sigma_j s_{i,j}\theta_j \geq 0\ ,\ i = 1, \ldots, m$ and $\Sigma_j(\Sigma_i s_{i,j})\theta_j = 1\}$. Consequently, the only densities we shall use will be the densities of the weights vector $\boldsymbol{\alpha}$, whose support can be chosen as the full Δ^{m-1}.

3. Statistical Inference

3.1 Bayes Theorem

The posterior density of the weights vector $p(\boldsymbol{\alpha}|\mathbf{n})$ is just proportional to the product of the prior density $p(\boldsymbol{\alpha})$ and the likelihood (1.1) rewritten as a function of $\boldsymbol{\alpha}$. We give the details for the case of Dirichlet prior distributed $\boldsymbol{\alpha}$.

Theorem 3.1 The likelihood (1.1) for i.i.d. sampling from a finite distribution with unknown probability vector $\boldsymbol{\theta}$ and the filtered-variate Dirichlet prior distribution $\boldsymbol{\theta} \sim F_{\mathbf{G}}D(\mathbf{a})$ (2.3) yield the posterior filtered-variate distribution, $\boldsymbol{\theta} = \mathbf{G}\boldsymbol{\alpha}$ where $\boldsymbol{\alpha}$ has the posterior density,

$$p(\boldsymbol{\alpha}|\mathbf{n}) = \left(B(\mathbf{a})^{-1}\prod_{j=1}^{m}\alpha_j^{a_j-1}\right)\times\prod_{i=1}^{k}\left(\sum_{j=1}^{m}\alpha_j g_{ij}\right)^{n_i} \Big/\ \mathcal{R}(\mathbf{a}, \mathbf{G}^T, -\mathbf{n})\ . \tag{3.1}$$

The normalizing constant in the density is a special case of Carlson's (1971) symmetrized multiple hypergeometric function.

$$\mathcal{R}(\mathbf{a}, \mathbf{G}^T, -\mathbf{n}) = \mathcal{R}_{n+}(\mathbf{a}, \mathbf{G}^T, -\mathbf{n})\ . \tag{3.2}$$

It is merely the complete integral of the numerator in (3.1), a Dirichlet expectation of the likelihood. Thus, it is the Bayesian prior predictive probability of a sample sequence $\mathbf{x}$ having frequency counts $\mathbf{n}$, $n_+ = N$, for fixed N,

$$p(\mathbf{x}) = \int p(\boldsymbol{\alpha}) \prod_{1}^{k} [\theta_i(\boldsymbol{\alpha})]^{n_i} \, d\boldsymbol{\alpha}$$

$$= \mathcal{R}(\mathbf{a}, G^T, -\mathbf{n}) . \qquad (3.3)$$

This posterior distribution of $\boldsymbol{\theta} = G\boldsymbol{\alpha}$ is induced by a posterior distribution of the weights vector $\boldsymbol{\alpha}$ (3.1) generalizing the Dirichlet. We denote this **generalized Dirichlet distribution** for the weights $\boldsymbol{\alpha}$ by

$$\boldsymbol{\alpha} \,|\mathbf{n} \sim D(\mathbf{a}, G^T, \mathbf{n}) . \qquad (3.4)$$

(Note the difference in sign between the third parameter in the distribution (3.4) and the final parameter in Carlson's function.) The distributions $D(\mathbf{a}, \mathbf{B}, \mathbf{c})$ were studied in Dickey (1983) and applied to missing-data problems in Dickey, Jiang, and Kadane (1987). Computation of $\mathcal{R}(\mathbf{a}, G^T, -\mathbf{n})$ and the moments of $D(\mathbf{a}, G^T, \mathbf{n})$ are available, as discussed later.

We note that by a multinomial expansion of the product of sums in (3.1), one can represent the generalized Dirichlet distribution, the posterior distribution of α, as a mixture of Dirichlet distributions. However, we have not found this to be a fruitful line of thinking in smooth inference problems. Such a mixture of Dirichlets seems useful purely as a technical device for computing integrals in cases of small or moderate n_+.

By Theorem 3.1, the posterior distribution of the linear transform $\boldsymbol{\theta} = G\boldsymbol{\alpha}$ is a **filtered-variate generalized Dirichlet distribution** since the posterior distribution of the weights $\boldsymbol{\alpha}$ is the generalized Dirichlet, $D(\mathbf{a}, G^T, \mathbf{n})$. We denote this posterior distribution of the sampling probabilities $\boldsymbol{\theta}$ by

$$\boldsymbol{\theta} \,|\mathbf{n} \sim F_G \, D(\mathbf{a}, G^T, \mathbf{n}) . \qquad (3.5)$$

A family of such distributions is obviously closed under further sampling from the same sampling distribution: a prior distribution $\boldsymbol{\theta} \sim F_G \, D(\mathbf{a}, G^T, \mathbf{n}^*)$ and sample data $\mathbf{n}$ would yield the posterior distribution $\boldsymbol{\theta} \,|\mathbf{n} \sim F_G \, D(\mathbf{a}, G^T, \mathbf{n}^* + \mathbf{n})$. An even greater generality, however, is available without any loss of tractability.

Define the **four-parameter filtered-variate Dirichlet distribution**, denoted by

$$\boldsymbol{\theta} \sim F_{\mathbf{G}} D(\mathbf{a}, \mathbf{H}^T, \mathbf{d}) , \tag{3.6}$$

as the distribution of $\boldsymbol{\theta} = \mathbf{G}\boldsymbol{\alpha}$, where $\boldsymbol{\alpha} \sim D(\mathbf{a}, \mathbf{H}^T, \mathbf{d})$.

Theorem 3.2 The four-parameter family of filtered-variate Dirichlet distributions is closed under sampling. If $\boldsymbol{\theta} \sim F_{\mathbf{G}} D(\mathbf{a}, \mathbf{H}^T, \mathbf{d})$ prior to ones knowledge of data $\mathbf{n}$ with likelihood (1.1), then ones coherent posterior distribution following $\mathbf{n}$ is

$$\boldsymbol{\theta} \,|\mathbf{n} \sim F_{\mathbf{G}} D[\mathbf{a}, (\mathbf{H}^T,\mathbf{G}^T)^T, (\mathbf{d}^T,\mathbf{n}^T)^T] . \tag{3.7}$$

Furthermore, if $\mathbf{H}^T = (\mathbf{K}^T, \mathbf{G}^T)$ and, conformably, $\mathbf{d}^T = (\mathbf{e}^T, \mathbf{n}^{*T})$, then

$$\boldsymbol{\theta} \,|\mathbf{n} \sim F_{\mathbf{G}} D[\mathbf{a}, (\mathbf{K}^T,\mathbf{G}^T)^T, (\mathbf{e}^T,(\mathbf{n}^*+\mathbf{n})^T)^T] . \tag{3.8}$$

3.2 Posterior Distribution of the Weights

We give further details, in this section, regarding the posterior distribution of the weights $\boldsymbol{\alpha}$ in the case of a Dirichlet prior for $\boldsymbol{\alpha}$. This generalized Dirichlet distribution, (3.1), (3.4), is tractable in several senses.

The moments of the linear filter $\boldsymbol{\theta} = \mathbf{G}\boldsymbol{\alpha}$ are easily calculated in terms of the moments of $\boldsymbol{\alpha}$. For $\mathbf{c} = (c_1, \ldots, c_m)^T$, the posterior $\mathbf{c}$th mixed moment of $\boldsymbol{\alpha}$, from (3.1), is proportional to a ratio of Carlson functions,

$$E\left(\prod_1^m \alpha_j^{c_j} \,|\mathbf{n}\right) = h(\mathbf{c}; \mathbf{a}) \times \mathcal{R}(\mathbf{a} + \mathbf{c}, \mathbf{G}^T, -\mathbf{n}) / \mathcal{R}(\mathbf{a}, \mathbf{G}^T, -\mathbf{n}) . \tag{3.9}$$

The proportionality factor h in (3.9) is just the corresponding prior $\mathbf{c}$th moment of $\boldsymbol{\alpha}$, as defined by (1.4).

Carlson's functions can easily be calculated by microcomputer, either as multinomial expansions, for small to medium N, or by Laplace's asymptotic method, for medium to large N (Tierney and Kadane 1986 and Dickey, Jiang, and Kadane 1987). The Laplace method will involve here the maximization of well behaved integrands that yield readily to a variety of optimization algorithms, including the E-M algorithm.

Results follow on the marginal and conditional distributions of the generalized Dirichlet. We exploit a lemma regarding the usual Dirichlet distribution, a distribution remarkable for the simplicity of its marginals and conditionals.

Lemma 3.3 (Adapted from Wilks 1962, pp 180-181). Conformably partition the vectors $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_{(1)}^T, \boldsymbol{\alpha}_{(2)}^T)^T$, $\mathbf{a} = (\mathbf{a}_{(1)}^T, \mathbf{a}_{(2)}^T)^T$, and define

$$\boldsymbol{\beta} = (\boldsymbol{\alpha}_{(1)}^T, \alpha_{(2)+})^T, \quad \boldsymbol{\gamma} = \boldsymbol{\alpha}_{(2)} / \alpha_{(2)+}, \tag{3.10}$$

where $\alpha_{(2)+}$ is the sum of the components of the vector $\boldsymbol{\alpha}_{(2)}$.
If $\boldsymbol{\alpha} \sim D(\mathbf{a})$, then $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ have the conditional and marginal Dirichlet distributions, respectively,

$$\boldsymbol{\gamma} \mid \boldsymbol{\beta} \sim D(\mathbf{a}_{(2)}), \qquad \boldsymbol{\beta} \sim D(\mathbf{b}), \tag{3.11}$$

where $\mathbf{b} = (\mathbf{a}_{(1)}^T, a_{(2)+})^T$.

Note that $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ are independent in the Dirichlet case.

Theorem 3.4 Define the vectors $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ in terms of $\boldsymbol{\alpha}$ as in the Lemma, eq. (3.10), but let $\boldsymbol{\alpha}$ have the generalized Dirichlet distribution, $\boldsymbol{\alpha} \sim D(\mathbf{a}, \mathbf{G}^T, \mathbf{n})$ (3.1). Conformably partition $\mathbf{G} = (\mathbf{G}_{(1)}, \mathbf{G}_{(2)})$ and $\boldsymbol{\alpha}$, $\mathbf{a}$ as in the Lemma. Then $\boldsymbol{\gamma}$ has the **conditional distribution** given $\boldsymbol{\beta}$,

$$\boldsymbol{\gamma} \mid \boldsymbol{\beta} \sim D[\mathbf{a}_{(2)}, \tilde{\mathbf{G}}_{(2)}(\boldsymbol{\beta})^T, \mathbf{n}], \tag{3.12}$$

where the matrix

$$\tilde{\mathbf{G}}_{(2)}(\boldsymbol{\beta}) = (\mathbf{G}_{(1)} \boldsymbol{\alpha}_{(1)}) \mathbf{1}_{m_2}^T + \mathbf{G}_{(2)} \alpha_{(2)+}. \tag{3.13}$$

Marginally, $\boldsymbol{\beta}$ has the density,

$$p(\boldsymbol{\beta}) = B(\mathbf{b})^{-1} \left(\prod_1^{m_1} \beta_j^{a_j - 1} \right) \beta_{m_1+1}^{a_{(2)+}-1} \, \mathcal{R}[\mathbf{a}_{(2)}, \tilde{\mathbf{G}}_{(2)}(\boldsymbol{\beta})^T, -\mathbf{n}] \\ / \, \mathcal{R}(\mathbf{a}, \mathbf{G}^T, -\mathbf{n}) , \qquad (3.14)$$

where, as in the Lemma, $\mathbf{b} = (\mathbf{a}_{(1)}^T, a_{(2)+})^T$.

Proof. We will distinguish the simpler distribution of the Lemma by use of an asterisk. Thus, we can refer to the simpler distribution as if $\boldsymbol{\alpha} \sim^* D(\mathbf{a})$, with density $p^*(\boldsymbol{\alpha})$. Then the density actually given in the theorem for $\boldsymbol{\alpha}$ satisfies

$$p(\boldsymbol{\alpha}) = p^*(\boldsymbol{\alpha}) \prod_{i=1}^k [\, \mathbf{g}_{i(1)}\boldsymbol{\alpha}_{(1)} + \mathbf{g}_{i(2)}\boldsymbol{\alpha}_{(2)} \,]^{n_i} \\ / \, \mathcal{R}(\mathbf{a}, \mathbf{G}^T, -\mathbf{n}) , \qquad (3.15)$$

where the row vectors of $\mathbf{G}_{(1)}$ and $\mathbf{G}_{(2)}$ are denoted by $\mathbf{g}_{i(1)}$ and $\mathbf{g}_{i(2)}$, respectively ($i = 1, \ldots, k$). Changing variables, we have since $\mathbf{1}_{m_2}^T \boldsymbol{\gamma} = 1$,

$$p_{\boldsymbol{\beta},\boldsymbol{\gamma}}(\boldsymbol{\beta},\boldsymbol{\gamma}) = p^*_{\boldsymbol{\beta},\boldsymbol{\gamma}}(\boldsymbol{\beta},\boldsymbol{\gamma}) \prod_i [\, \mathbf{g}_{i(1)}\boldsymbol{\alpha}_{(1)}\mathbf{1}_{m_2}^T\boldsymbol{\gamma} + \mathbf{g}_{i(2)}\boldsymbol{\alpha}_{(2)+}\boldsymbol{\gamma} \,]^{n_i} \\ / \, \mathcal{R}(\mathbf{a}, \mathbf{G}^T, -\mathbf{n}) , \qquad (3.16)$$

Note that the Jacobian is contained in the first factor $p^*_{\boldsymbol{\beta},\boldsymbol{\gamma}}$. Since $\boldsymbol{\gamma} \,|\, \boldsymbol{\beta} \sim^* D(\mathbf{a}_{(2)})$ and $\boldsymbol{\beta} \sim^* D(\mathbf{b})$, then

$$p_{\gamma|\beta}(\gamma|\beta)\, p_{\beta}(\beta) =$$

$$p^*_{\gamma|\beta}(\gamma|\beta) \prod_i [\, \tilde{g}_{i(2)}(\beta)\, \gamma \,]^{n_i} \;/\; \mathcal{R}[a_{(2)}, \tilde{G}_{(2)}(\beta)^T, -n]$$

$$p^*_{\beta}(\beta)\; \mathcal{R}[a_{(2)}, \tilde{G}_{(2)}(\beta)^T, -n] \;/\; \mathcal{R}(a, G^T, -n)\,, \qquad (3.17)$$

3.3 Posterior Estimates

The posterior mean is a convenient and theoretically attractive estimate to use for θ. The mean of a filtered-variate distribution $\theta = G\alpha$ is the same linear filter of the mean of the underlying vector α,

$$E(\theta|n) = G\, E(\alpha|n)\,, \qquad (3.18)$$

and the posterior variance matrix is quadratic in the posterior variance matrix of α,

$$\mathrm{Var}(\theta|n) = G\, \mathrm{Var}(\alpha|n)\, G^T\,. \qquad (3.19)$$

In the prior filtered-variate Dirichlet case, as we have seen, the posterior moments of the weights α are ratios of Carlson functions (3.9).

The linear relation for the posterior mean (3.18) depends in no way on the rank of the filter matrix G. When G is nonsingular the posterior mode of θ, too, is conveniently available as the linear filter of the posterior mode of α.

Theorem 3.5 If G is nonsingular and $\theta = G\alpha$, then

$$\mathrm{Mode}(\theta|n) = G\, \mathrm{Mode}(\alpha|n)\,. \qquad (3.20)$$

Proof. For nonsingular G, the transformation is one-to-one, the Jacobian is constant, and so the posterior densities of θ and α are directly proportional (as in the prior (2.7)).

To appreciate that nonsingularity of **G** is important for the invariance of the mode (3.20), note that unlike the mean, a mode is not preserved under marginalization. For example, in the case of a Dirichlet distribution, $(\alpha_1, \ldots, \alpha_k) \sim D(a, b, \ldots, b)$, we have that $\text{Mode}(\alpha_1, \ldots, \alpha_{k-1}) = (a-1, b-1, \ldots, b-1) / [a + (k-1)b - k]$, by (1.10), but since $\alpha_1 \sim \text{Beta}[a, (k-1)b]$, we have $\text{Mode}(\alpha_1) = (a-1) / [a + (k-1)b - 2]$.

3.4 Comparing Hypotheses

Posterior "scientific reporting" was defined in Dickey (1973) to require graphical or similar display of the dependence of the inference on prior distributions meaningfully interpreted in the real-world problem. Bayesian comparative judgement of hypotheses is based on the posterior odds for one hypothesis versus another, $P(H_1 \mid \text{Data})/P(H_2 \mid \text{Data})$. The evidence in statistical data relevant to such a judgement is summarized through the Bayes factor, the ratio of the posterior odds to the prior odds. The Bayes factor can be calculated from the data and the prior distributions (conditional on each of the two hypotheses) as the ratio of the two predictive probabilities (or densities) of the data,

$$[P(H_1 \mid \text{Data})/P(H_2 \mid \text{Data})] \,/\, [P(H_1)/P(H_2)]$$

$$= p(\text{Data} \mid H_1) \,/\, p(\text{Data} \mid H_2) \,.$$

In general, a predictive probability is the integral function of the prior distribution $P(\theta \mid H)$ under an hypothesis H, $p(\text{Data} \mid H) = \int p(\text{Data} \mid \theta)\, dP(\theta \mid H)$. In Theorem 3.1, (3.3), we obtained the predictive probability of a sample sequence, Data = $\mathbf{x} = (x_1, \ldots, x_N)^T$, for a filtered-variate Dirichlet prior distribution. An ordinary Dirichlet prior distribution would have the predictive probability $h(\mathbf{n}; \mathbf{b})$ (1.4). Thus, our Bayes factor for comparing histogram sampling models, at least one of which is locally smooth, is a simple ratio involving one or more Carlson functions.

4. Distributions with Mean-Structured Variance

Several methodologies can be considered for prior assessment of local smoothness. We will propose new methods involving judgements of typical configurations and prior moments. It will be useful, first, to present a general theory of distributions on the probability simplex having a structured variance matrix depending quadratically on the mean vector. (Related properties in one dimension were studied by Bar-Lev and Enis 1986.) This is followed by a second-order representation theory for filtered variates of such distributions.

Lemma 4.1 Suppose a random vector $\mathbf{y} = (y_1, \ldots, y_k)^T$ has the first two moments,

$$E\,\mathbf{y} = \boldsymbol{\mu} \ , \tag{4.1a}$$

$$\mathrm{Var}(\mathbf{y}) = c\,[\mathbf{Diag}(\boldsymbol{\mu}) - \boldsymbol{\mu}\boldsymbol{\mu}^T] \ , \tag{4.1b}$$

for some fixed vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)^T$ and some scalar $c \geq 0$. If $\mu_+ =$ 0, or 1, respectively, then $y_+ = 0$ w.p.1, or 1 w.p.1.

Proof. $E(y_+) = \mu_+$ and $\mathrm{Var}(y_+) = c\,\mu_+(1-\mu_+)$.

Our interest is in random probability vectors. So if $\mathbf{y}$ has the moments (4.1) with $\mu_+ = 1$, we will say that $\mathbf{y}$ has a distribution with **mean-structured variance (MSV)** and write

$$\mathbf{y} \sim \mathrm{MSV}(\boldsymbol{\mu}, c) . \tag{4.2}$$

The normalized multinomial, for example, and the Dirichlet are MSV distributions.

Lemma 4.2 If $\mathbf{n} \sim$ multinomial$(N, \boldsymbol{\theta})$, then for $\mathbf{y} = \mathbf{n}/N$, $\mathbf{y} \sim$ $\mathrm{MSV}(\boldsymbol{\theta}, 1/N)$. If $\mathbf{y} \sim D(\mathbf{b})$, with $\mathbf{b} = \mathbf{w}b_+$, then $\mathbf{y} \sim \mathrm{MSV}[\mathbf{w}, 1/(b_+ + 1)]$.

A prominent feature of the MSV property is its preservation under arbitrarily weighted averaging of independent vectors having the same mean.

Theorem 4.3 If $\mathbf{y}_i \sim \text{MSV}(\boldsymbol{\mu}, c_i)$, $i = 1, 2$, are uncorrelated, $v_i \geq 0$, $i = 1, 2$, and $v_1 + v_2 = 1$, then

$$v_1\mathbf{y}_1 + v_2\mathbf{y}_2 \sim \text{MSV}(\boldsymbol{\mu},\ v_1^2 c_1 + v_2^2 c_2). \tag{4.3}$$

The marginalization or grouping behavior of the multinomial distribution and the Dirichlet distribution (*Lemma 3.3*) extends to all MSV distributions. The MSV structure is preserved under grouping.

Theorem 4.4 Let $\{S(1), \ldots, S(h)\}$ be a partition of $\{1, \ldots, k\}$. If $\mathbf{y} \sim \text{MSV}(\boldsymbol{\mu}, c)$, then $\mathbf{z} \sim \text{MSV}(\boldsymbol{\zeta}, c)$, where $\mathbf{z}$ and $\boldsymbol{\zeta}$ have respective coordinates $z_i = \Sigma_{S(i)}\, y_j$ and $\zeta_i = \Sigma_{S(i)}\, \mu_j$, for $i = 1, \ldots, h$.

4.1 Mean-Mixing and Scale-Mixing.

We shall say that the distribution of a random vector $\mathbf{y}$ is a **mean-mixture** or **mean-compound** of the conditional distributions, $\mathbf{y} \mid \mathbf{x} \sim \text{MSV}(\mathbf{x}, c)$, by the distribution of the random vector $\mathbf{x}$ iff c does not depend on $\mathbf{x}$. In this case, $E(\mathbf{y}) = E(\mathbf{x})$. For example, if $(\mathbf{n} \mid \boldsymbol{\theta}) \sim \text{multinomial}(N, \boldsymbol{\theta})$, and $\boldsymbol{\theta} \sim D(\mathbf{b})$, then the marginal distribution of $\mathbf{n}/N$ is the mean-mixture of a normalized multinomial distribution by a Dirichlet distribution. The marginal distribution of such $\mathbf{n}$ is known as the Dirichlet-multinomial$(N, \mathbf{b})$, the usual conjugate Bayesian prior predictive distribution for multinomial sampling. All three distributions in this example are MSV. We find, in general, that the MSV property is preserved under mean-mixing of one MSV distribution by another MSV distribution.

Theorem 4.5 If $\mathbf{y} \mid \mathbf{x} \sim \text{MSV}(\mathbf{x}, c)$ and $\mathbf{x} \sim \text{MSV}(\boldsymbol{\mu}, d)$, then $\mathbf{y} \sim \text{MSV}(\boldsymbol{\mu}, c+d-cd)$.

Proof. Merely write out $\text{Var}(\mathbf{y}) = E[\text{Var}(\mathbf{y} \mid \mathbf{x})] + \text{Var}[E(\mathbf{y} \mid \mathbf{x})]$ and replace $E(\mathbf{x}\mathbf{x}^T)$ by $\text{Var}(\mathbf{x}) + (E\mathbf{x})(E\mathbf{x})^T$.

Corollary 4.6 If $\mathbf{n} \sim \text{Dirichlet-multinomial}(N, \mathbf{b})$, where $b = \mathbf{w}b_+$, then for $\mathbf{y} = \mathbf{n}/N$, $\mathbf{y} \sim \text{MSV}\{\mathbf{w}, (N+b_+)/[N(b_+ + 1)]\}$.

Similar corollaries can be stated immediately for other compounds of MSC distributions, such as normalized multinomial by normalized multinomial, Dirichlet by Dirichlet, and even the bazaar mean compound of Dirichlet by normalized multinomial (in the stated order).

Not surprisingly, we also find that the MSV property is preserved under arbitrary scale mixing.

Theorem 4.7 If $\mathbf{y} \mid \mathbf{x} \sim \mathrm{MSV}(\boldsymbol{\mu}, c)$ and $\boldsymbol{\mu}$ does not depend on c, then $\mathbf{y} \sim \mathrm{MSV}(\boldsymbol{\mu}, E(c))$.

4.3 Filtered-Variate MSV Distributions and Representations by Extreme Distributions on a Convex Polytope

MSV distributions have a simple limiting form that will be useful in the assessment of filtered-variate prior distributions.

Theorem 4.8 For $\mathbf{z} = (z_1, \ldots, z_m)^T$, if $\mathbf{z} \sim \mathrm{MSV}(\mathbf{u}, c)$ then $c \leq 1$. Furthermore, as $c \to 1$, $\mathbf{z}$ has the finite limiting distribution $\tilde{P}$ supported on the vertices of the probability simplex Δ^{m-1},

$$\tilde{P}[\mathbf{z} = \boldsymbol{\delta}_{(j)}] = u_j, \tag{4.7}$$

for $j = 1, \ldots, m$, where $\delta_{(j)j} = 1$ and $\delta_{(j)i} = 0$ for $i \neq j$.

Lemma 4.9 If the (scalar) random quantity z is supported on the unit interval, $0 \leq z \leq 1$, and $Ez = u$, then the variance of z is maximized for fixed u by $P(z = 1) = u$, $P(z = 0) = 1 - u$.

Proof of Lemma. By a standard linear optimization argument, the extreme measure under two linear constraints (total measure unity and mean fixed) is a two-point distribution, $P(z = z_{(j)}) = p_j$, $j = 1, 2$. Then $\mathrm{Var}(z) = p_1 p_2 (z_{(1)} - z_{(2)})^2 = (z_{(2)} - u)(u - z_{(1)})$, which is maximized by $z_{(1)} = 0$, $z_{(2)} = 1$.

Corollary 4.10 If $\mathbf{z} \sim D(\mathbf{a})$, $\mathbf{a} = \mathbf{u} a_+$, and $a_+ \to 0$ [$c = (a_+ + 1)^{-1} \to 1$], the limiting distribution of $\mathbf{z}$ is the finite distribution $\tilde{P}$ (4.7).

We turn again to the idea of filtering a random vector, recalling that for a filtered-variate Dirichlet, the underlying Dirichlet distribution is MSV. Consider the random vector

$$\boldsymbol{\theta} = \mathbf{G}\boldsymbol{\alpha}, \tag{4.8}$$

assuming $\boldsymbol{\alpha} \sim \text{MSV}(\mathbf{u}, c)$. We shall say that the induced distribution of $\boldsymbol{\theta}$ is **filtered-variate mean-structured variance** (**FMSV**), and write $\boldsymbol{\theta} \sim F_{\mathbf{G}}\text{MSV}(\mathbf{u}, c)$. Again, $\boldsymbol{\theta}$ has a distribution on the convex hull $H\!I(\mathbf{G})$ of the columns of $\mathbf{G} = (\mathbf{g}_1, \ldots, \mathbf{g}_m)$.

Our limiting form of an MSV distribution (4.7) implies a finitely supported limiting form for the FMSV, as follows. (This can also be considered as giving a representation for any finite distribution, whatsoever, as an extreme FMSV distribution.)

Theorem 4.11 For $\boldsymbol{\theta} \sim F_{\mathbf{G}}\text{MSV}(\mathbf{u}, c)$, as $c \to 1$, $\boldsymbol{\theta}$ has the limiting distribution $\tilde{P}_{\mathbf{G}}$ finitely supported on the set of column vectors of $\mathbf{G}$,

$$\tilde{P}_{\mathbf{G}}[\boldsymbol{\theta} = \mathbf{g}_j] = u_j, \tag{4.9}$$

for $j = 1, \ldots, m$.

The finite support of $\tilde{P}_{\mathbf{G}}$, the set of the column vectors of $\mathbf{G}$, includes as a subset the extreme points or vertex points of $H\!I(\mathbf{G})$.

We obtain an important and interesting second-order representation of an FMSV distribution in terms of the finite limiting distribution, from the moment relations, $E\boldsymbol{\theta} = \mathbf{G}(E\boldsymbol{\alpha})$, $\text{Var}(\boldsymbol{\theta}) = \mathbf{G}\,\text{Var}(\boldsymbol{\alpha})\mathbf{G}^T$, and the MSV property of $\boldsymbol{\alpha}$.

Corollary 4.12 If $\boldsymbol{\theta} \sim F_{\mathbf{G}}\text{MSV}(\mathbf{u}, c)$, then

$$E\,\boldsymbol{\theta} = E\,\tilde{\boldsymbol{\theta}} = \Sigma\, u_j\, \mathbf{g}_j, \tag{4.10}$$

and

$$\begin{aligned} \text{Var}(\boldsymbol{\theta}) &= c\,\text{Var}(\tilde{\boldsymbol{\theta}}) \\ &= c\,\Sigma\, u_j(\mathbf{g}_j - E\tilde{\boldsymbol{\theta}})(\mathbf{g}_j - E\tilde{\boldsymbol{\theta}})^T, \end{aligned} \tag{4.11}$$

where $\tilde{\boldsymbol{\theta}}$ has the finite distribution $\tilde{P}_{\mathbf{G}}$ (4.9) supported on the set of column vectors of $\mathbf{G}$.

Corollary 4.13 If $\boldsymbol{\theta} \sim F_{\mathbf{G}}MSV(\mathbf{u}, c)$, $Corr(\boldsymbol{\theta}) = Corr(\tilde{\boldsymbol{\theta}})$. (4.12)

Corollary 4.14 Both the limiting distribution (4.9) and the moment representations (4.10), (4.11) hold for the filtered-variate Dirichlet, $\boldsymbol{\theta} \sim F_{\mathbf{G}}D(\mathbf{a})$, $\mathbf{a} = \mathbf{u}a_+$, as $a_+ \to 0$.

5. Prior Assessment

5.1 Prior-Typical Probability Vectors

A prior distribution $p(\boldsymbol{\theta})$ for unknown $\boldsymbol{\theta}$ allows statements that some vector values are more probable than others. Can $p(\boldsymbol{\theta})$ be used to give emphasis to vector values $\boldsymbol{\theta}$ that are "typical" of the prior uncertainty represented by $p(\boldsymbol{\theta})$? Both yes and no. "No" in the sense that natural summaries like the mean and mode of $p(\boldsymbol{\theta})$ can be quite untypical of random outcomes $\boldsymbol{\theta}$ from $p(\boldsymbol{\theta})$. Mere pairwise symmetry in which the coordinate components θ_i all have the same mean would imply the mean vector $E(\boldsymbol{\theta}) = (1/k, \ldots, 1/k)^T$. This is usually far too smooth to be a value of vector $\boldsymbol{\theta}$ typifying ones prior uncertainty. A further assumption of joint unimodality would imply this same point as the prior mode. For example, the vector of means of an i.i.d. normal sequence, which is also the mode, is quite untypical of vector outcomes of the distribution. This is still so if the joint normal distribution has moderate serial correlations. Similarly, for a symmetric Dirichlet distribution, $\boldsymbol{\theta} \sim D(b\mathbf{1}_k)$, $Mode(\boldsymbol{\theta}) = E(\boldsymbol{\theta}) = (1/k, \ldots, 1/k)^T$. So even the most probable or central vector value can fail to be typical of a given multivariate distribution.

What is wrong is that we have not yet chosen a scale or measure by which to judge the smoothness of random $\boldsymbol{\theta}$. If a smoothness summary function $sm(\boldsymbol{\theta})$, is specified, having large values of $sm(\boldsymbol{\theta})$ for vectors $\boldsymbol{\theta}$ considered "rough" and small values of $sm(\boldsymbol{\theta})$ for vectors considered "smooth," we can examine the induced distribution of the statistic $sm(\boldsymbol{\theta})$ to see what values of $sm(\boldsymbol{\theta})$ are probable and what values of $sm(\boldsymbol{\theta})$ are

rare, and hence, how much smoothness is typical of $\boldsymbol{\theta}$ and what is the range of prior variation in smoothness. The distribution of $sm(\boldsymbol{\theta})$ does not necessarily emphasize a region of $\boldsymbol{\theta}$ values that includes the mode or mean of $\boldsymbol{\theta}$. For example, the chi-squared statistic for an i.i.d. normal vector $\boldsymbol{\theta} \sim N^{(k)}(\mathbf{0}, \mathbf{I})$, $sm^2(\boldsymbol{\theta}) = \Sigma\, \theta_i^2 \sim \chi^2_k$, emphasizes vector values $\boldsymbol{\theta}$ for which $\left|(\Sigma\, \theta_i^2) - k\right| \leq 2\sqrt{(2k)}$.

Similar considerations arise in statistical physics when macrovariables are introduced, physically measurable functions of the microstate. "A meaningful macroworld description involves an averaging that washes out information of the microworld, and it is we who average. . . . it is we who impose the macroscopic description of physical reality, a reality which does not apply to the microworld." (Pagels 1982, p 109) See Jaynes (1979) on the use of entropy as a summary function to indicate typical values for a vector of frequencies.

We have particular interest in quadratic functions $sm^2\boldsymbol{\theta})$ that measure local smoothness. One obvious measure is a squared difference $(\theta_i - \theta_j)^2$. We assume in this paper that if a statistician reports an estimate for $\boldsymbol{\theta}$, he desires his estimate to be about as smooth as the true vector $\boldsymbol{\theta}$. Unfortunately, any unbiased estimate, such as $\hat{\boldsymbol{\theta}} = \mathbf{n}/n_+$, will be less smooth, in the sense of squared differences, than the true $\boldsymbol{\theta}$, according to the prior belief concerning $\boldsymbol{\theta}$,

$$E[(\hat{\theta}_i - \hat{\theta}_j)^2] = E[(\theta_i - \theta_j)^2] + E[\mathrm{Var}(\hat{\theta}_i - \hat{\theta}_j \,|\, \boldsymbol{\theta})]$$

$$\geq E[(\theta_i - \theta_j)^2]\,. \qquad (5.1)$$

The posterior distribution coherently combines the information regarding smoothness from the prior and from the data, and we suggest using the posterior mean or mode as a not untypically smooth representative of the posterior distribution. This contrasts with the prior mean or mode, which, because of symmeties in the prior distribution, is less typical of the smoothness in the prior than is the posterior mean or mode in regard to the smoothness in the posterior distribution.

If two category sampling probabilities have the same prior mean, then $\mathrm{Var}(\theta_i - \theta_j)$ can serve as a prior typical value of the squared difference $(\theta_i - \theta_j)^2$, and this can be computed from the prior variance matrix of $\boldsymbol{\theta}$, as

$$\mathrm{Var}(\theta_i - \theta_j) = [\mathrm{Var}(\theta_i) + \mathrm{Var}(\theta_j)] - 2\mathrm{Cov}(\theta_i, \theta_j). \quad (5.2)$$

So high prior covariances for neighboring categories will express a prior opinion that the unknown probability vector is locally smooth, provided that the mean differences are small. A variance matrix with large entries, but small variance of the difference (5.2), for i near j, would express highly vague uncertainty about $\boldsymbol{\theta}$, but strong prior prejudice in favor of $\boldsymbol{\theta}$ being locally smooth. Consider a j-distant squared difference $(\theta_i - \theta_{i+j})^2$. Under suitable approximate stationarity assumptions, the induced prior distribution of $\mathrm{sm}(\boldsymbol{\theta})$ does not depend much on the category i and can provide useful information regarding prior local smoothness. The local smoothness of the vector $\boldsymbol{\theta}$, as a whole, is measured by the root mean squared j-distant difference,

$$\mathrm{sm}^2(\boldsymbol{\theta}) = (k-j)^{-1} \sum_{i=1}^{k-j} (\theta_i - \theta_{i+j})^2 , \quad (5.3)$$

for which the expectation of $\mathrm{sm}^2(\boldsymbol{\theta})$, again, is straightforward from the prior mean and variance matrix of $\boldsymbol{\theta}$. Our methods for constructing prior distributions will emphasize the attainment of desired first and second order prior moments.

Two basic inference situations and prior-assessment goals are of particular interest. First, the expert who wants to take formal account of his full prior information will want a relatively accurately assessed prior mean vector and variance matrix and a distribution of $\mathrm{sm}(\boldsymbol{\theta})$ representing his prejudiced, but still uncertain, prior opinion concerning the local smoothness of $\boldsymbol{\theta}$. He will want the prior-support polytope $H\!/\!/(\mathbf{G})$ to have full dimensionality, k-1, appreciable volume, and to bear a well contoured prior density. Another expert may want to limit the prior input to a strong belief concerning the local smoothness of $\boldsymbol{\theta}$ with an otherwise highly diffuse, or "noninformative", opinion regarding $\boldsymbol{\theta}$. In the latter situation, $H\!/\!/(\mathbf{G})$ may contain a variety of probability vectors that are all locally smooth, with small $\mathrm{sm}(\boldsymbol{\theta})$ or missing high frequency components, but have a diffuse distribution of $\boldsymbol{\theta}$ within $H\!/\!/(\mathbf{G})$. The posterior estimate will then turn out to be much like a projection of $\hat{\boldsymbol{\theta}} = \mathbf{n}/n_+$ into $H\!/\!/(\mathbf{G})$.

5.2 Symmetric Priors

In order to construct a filtered-variate form of prior $F_{\mathbf{G}}MSV(\mathbf{u}, c)$ for $\boldsymbol{\theta}$ with a particular desired mean vector and variance matrix one might first develop a discrete (finite) distribution $\tilde{P}_{\mathbf{G}}$ for $\tilde{\boldsymbol{\theta}}$ having the desired means, $E\,\boldsymbol{\theta} = E\,\tilde{\boldsymbol{\theta}}$ (4.10), and correlations, $Corr(\boldsymbol{\theta}) = Corr(\tilde{\boldsymbol{\theta}})$ (4.12), and then arrange the remaining variance-proportionality c to achieve the desired variances, $Var(\boldsymbol{\theta}) = c\,Var(\tilde{\boldsymbol{\theta}})$, $Var(\boldsymbol{\theta}) = c\,Var(\tilde{\boldsymbol{\theta}})$. For simplicity, it is tempting to begin by examining the case where the discretely distributed $\tilde{\boldsymbol{\theta}}$ is uniform over its finite support set,

$$P[\tilde{\boldsymbol{\theta}} = \mathbf{g}_j] = u_j = 1/m\,, \tag{5.4}$$

for $j = 1, \ldots, m$. For a one-dimensional histogram with sequentially numbered categories, consider a square matrix $\mathbf{G}$, $k \times m$ with $m = k$, where each point $\mathbf{g}_j$ is a j-shifted version of a single basic probability vector $\mathbf{t} = (t_1, \ldots, t_k)^T$,

$$g_{ij} = t_r\,, \qquad r = (i-j+1) \bmod k. \tag{5.5}$$

This would imply an ergodic property for $\tilde{P}_{\mathbf{G}}$, whereby the joint distribution of a set of coordinates of $\tilde{\boldsymbol{\theta}}$ would be identical to their serial distribution over the category shifts of any fixed realization $\mathbf{g}_j$. Then $E\,\theta_i = E\,\tilde{\theta}_i \equiv \bar{t} = 1/k$, and similarly, the variance would be a Toeplitz-form matrix, $\sigma_{i,i+j} = \tilde{\sigma}_{i,i+j} \equiv k^{-1}\Sigma_r\, t_r t_{r+j} = f(j)$. Such cyclical translation invariance in one or more dimensions provides a powerful simplicity, but it may need to be corrected, in use, for unrealistic edge effects.

Somewhat less simply, one can increase m beyond k ($m > k$) and include different shapes or "frequencies" along with the location shifts or "phases", e.g.

$$t_{r,r'} \propto 1 + \sin[(rB_{r'} - C_{r'})2\pi/k]\,. \tag{5.6}$$

Such a frequency-and-phase approach is related to that of Lo (1984) in the continuous realm.

Once a distribution for $\tilde{\boldsymbol{\theta}}$, in the form of extreme vectors $\mathbf{G} = (\mathbf{g}_1, \ldots, \mathbf{g}_m)$ and weights $\mathbf{u}$, has been chosen, one can assess a prior distribution for $\boldsymbol{\theta}$ as $\boldsymbol{\theta} \sim F_{\mathbf{G}}\mathrm{MSV}(\mathbf{u}, c)$ where c is chosen to give the desired ratio of variances, $\mathrm{Var}(\boldsymbol{\theta}) = c\,\mathrm{Var}(\tilde{\boldsymbol{\theta}})$. A filtered-variate Dirichlet prior distribution, in particular, would then be assessed as

$$\boldsymbol{\theta} \sim F_{\mathbf{G}}D[(c^{-1} - 1)\mathbf{u}]. \tag{5.7}$$

One way the prior variance of $\boldsymbol{\theta}$, needed to determine c, could be obtained is by eliciting expert opinion in the form of a subjectively typical set of $\boldsymbol{\theta}$-vectors, the $\mathbf{u}$-weighted empirical moments of which could then be calculated. An approximate matrix proportionality, $\mathrm{Var}(\boldsymbol{\theta}) = c\,\mathrm{Var}(\tilde{\boldsymbol{\theta}})$, may be difficult to achieve by an abstract mathematical choice of columns for $\mathbf{G}$. An alternative way for choosing $\mathbf{G}$ will be considered.

5.3 Using Elicited Typical Vectors

Because $\mathrm{Var}(\boldsymbol{\theta}) = c\,\mathrm{Var}(\tilde{\boldsymbol{\theta}})$ and because the inequality $0 < c < 1$ is required for $\boldsymbol{\theta} \sim F_{\mathbf{G}}\mathrm{MSV}(\mathbf{u}, c)$ to be continuously distributed (Theorems 4.7, 4.10), the discrete distribution for $\tilde{\boldsymbol{\theta}}$ must be given variances and covariances that are too large for the discrete distribution, itself, to model the desired prior smoothness directly. Hence, the points of support $\mathbf{g}_j$ for $\tilde{\boldsymbol{\theta}}$ cannot, themselves, be typical probability vectors chosen for their typical smoothness. Yet, they can be developed from such typical vectors by use of the following result, in which we increase the variability of a set of probability vectors by subtracting from each vector a fraction of the mean vector and then renormalizing each vector. Denote by $\tilde{\boldsymbol{\theta}}_{\mathbf{H}}$ a random vector having the finite distribution over the set of column vectors of the matrix $\mathbf{H} = (\mathbf{h}_1, \ldots, \mathbf{h}_m)$,

$$P[\tilde{\boldsymbol{\theta}}_{\mathbf{H}} = \mathbf{h}_j] = u_j, \tag{5.8}$$

for $j = 1, \ldots, m$. The low-order moments of $\tilde{\boldsymbol{\theta}}_{\mathbf{H}}$ are $E\,\tilde{\boldsymbol{\theta}}_{\mathbf{H}} = \Sigma\, u_j \mathbf{h}_j = \overline{\mathbf{h}}$ and $\mathrm{Var}(\tilde{\boldsymbol{\theta}}_{\mathbf{H}}) = \Sigma\, u_j(\mathbf{h}_j - \overline{\mathbf{h}})(\mathbf{h}_j - \overline{\mathbf{h}})^T$.

Theorem 5.1 Assume that all the columns of the matrix $\mathbf{H}$ are probability vectors and all the coordinates of the mean vector $\overline{\mathbf{h}}$ are positive. Define for given $-\infty < \delta \le \min(h_{j,i}/\overline{h}_i)$,

$$\mathbf{G} = (1-\delta)^{-1}(\mathbf{H} - \delta\overline{\mathbf{h}}\,\mathbf{1}^T). \tag{5.9}$$

Then all the columns of $\mathbf{G} = (\mathbf{g}_1, \ldots, \mathbf{g}_m)$ are probability vectors, and the random vector $\tilde{\boldsymbol{\theta}}_{\mathbf{G}}$, where $P[\tilde{\boldsymbol{\theta}}_{\mathbf{G}} = \mathbf{g}_j] = u_j$, has the moments,

$$E\,\tilde{\boldsymbol{\theta}}_{\mathbf{G}} = E\,\tilde{\boldsymbol{\theta}}_{\mathbf{H}} \qquad \text{and} \qquad \mathrm{Var}(\tilde{\boldsymbol{\theta}}_{\mathbf{G}}) = (1-\delta)^{-2}\,\mathrm{Var}(\tilde{\boldsymbol{\theta}}_{\mathbf{H}}). \tag{5.10}$$

Furthermore, if $\delta \ge 0$, the convex hull $H\!I(\mathbf{G})$ contains the set of column vectors of $\mathbf{H}$.

Proof. To prove the final assertion of the theorem, note that since $\overline{\mathbf{h}} = \overline{\mathbf{g}}$, each $\mathbf{h}_j = (1-\delta)\mathbf{g}_j + \delta\,\overline{\mathbf{g}}$.

A tentative method for assessing an FMSV prior distribution proceeds by the following three steps. Again, for simplicity, we work with equal weights, each $u_j = 1/m$, $j = 1, \ldots, m$.

1. Elicit from the expert a set of $m \ge k$ typical category-sampling probability vectors, $\boldsymbol{\theta} = \mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_m$. These are chosen to be typical for their local smoothness and to have an empirical mean vector and variance matrix matching the expert's personal means and variances of vector $\boldsymbol{\theta}$,

$$E\,\boldsymbol{\theta} = m^{-1}\Sigma\,\mathbf{h}_j = \overline{\mathbf{h}} \tag{5.11}$$

$$\mathrm{Var}(\boldsymbol{\theta}) = m^{-1}\Sigma\,(\mathbf{h}_j - \overline{\mathbf{h}})(\mathbf{h}_j - \overline{\mathbf{h}})^T. \tag{5.12}$$

2. For $\mathbf{H} = (\mathbf{h}_1, \ldots, \mathbf{h}_m)$, define the matrix $\mathbf{G}$ according to (5.9) and let $\boldsymbol{\theta} \sim F_{\mathbf{G}}\mathrm{MSV}[m^{-1}\mathbf{1}_m, (1-\delta)^2]$, a class of distributions having the same moments, (5.11), (5.12), as the expert by Corollary 4.12 and (5.10). In the case of a filtered-variate Dirichlet prior distribution we obtain

$$\boldsymbol{\theta} \sim F_{\mathbf{G}}D(a\mathbf{1}_m), \tag{5.13}$$

where, by Lemma 4.2, $(ma + 1)^{-1} = (1 - \delta)^2$, that is, $a = a(\delta)$ where

$$a(\delta) = [(1-\delta)^{-2} - 1] / m \ . \qquad (5.14)$$

3. Fine-tune δ to model the expert's prior range of uncertainty regarding the smoothness of $\boldsymbol{\theta}$. This can be done by considering the induced distribution of a smoothness summary $sm(\boldsymbol{\theta})$, or by showing to the expert monte carlo samples from (5.12).

Two practical problems arise immediately. First, the expert may find it more difficult to devise a list of vectors having, as empirical moments, his preconceived overall prior means and variances than just to report several vectors that are typically smooth. Secondly, enough vectors are needed to generate a sufficiently rich convex hull. Our final theorem will suggest that both these problems can be handled by a method combined from the methods of Sections 5.2 and 5.3. If the expert avows a symmetry like (5.5), or for some other reason, would like to express a prior distribution in which $E\theta_i \equiv 1/k$ and $Cov(\theta_i, \theta_{i+j}) \equiv f(j)$, these moments can be achieved by construction, by extending his elicited list of typical vectors to include all their category shifts. Or, if a nonconstant mean $E\theta_i$ is desired, the shifts can be performed on the difference vectors, typical vectors minus desired mean vector. As mentioned in Section 5.1, the mean, itself, should not be considered a typical vector.

For clarity of exposition, only, we state the following results in terms of a single typical vector $\mathbf{h}_0$ and the desired mean vector $\mathbf{p}$.

Theorem 5.2 Given two probability vectors, $\mathbf{h}_0$ and $\mathbf{p}$, define $\mathbf{t}^*$ by

$$\mathbf{h}_0 = \mathbf{p} + \mathbf{t}^* \ , \qquad (5.15)$$

and define the shifted difference vectors of $\mathbf{t}^*$ according to (5.5), $\mathbf{g_j}^* = (g_{1j}^*, \ldots, g_{kj}^*)^T$, $j = 1, \ldots, k$. (Here, $\Sigma\, t_i^* = 0$, and similarly for each $\mathbf{g}_j^*$.) Then if $\min(t_i^*) + \min(p_i) \geq 0$, the vectors in the new list,

$$\mathbf{h}_j = \mathbf{p} + \mathbf{g}_j^* \ , \qquad (5.16)$$

$j = 1, \ldots, k$, are probability vectors, and the new list has the empirical mean, $\overline{\mathbf{h}} = k^{-1}\Sigma\, \mathbf{h}_j = \mathbf{p}$, and an empirical variance matrix in Toeplitz

form matching the serial covariances from $\mathbf{t}^*$, $\tilde{\sigma}_{i,i+j} = k^{-1}\Sigma_i t_i t_{i+j} = f(j)$.

Again, in the constant-mean case, the shifts can be applied directly to the typical vectors, themselves, since the shifts would have no effect on the mean vector. Finally, in this case, if $\boldsymbol{\theta} \sim F_{\mathbf{G}}MSV[k^{-1}\mathbf{1}_k, (1-\delta)^2]$, say, where $m = k$ and $\mathbf{G}$ is derived by (5.9) from the shifts $\mathbf{H}$ on a single typical vector, then the prior expectation of the j-distant local smoothness summary, $sm^2(\boldsymbol{\theta}) = (\theta_i - \theta_{i+j})^2$, will be the same as the empirical serial smoothness of the expert's typical vector,

$$E\, sm^2(\boldsymbol{\theta}) = k^{-1}\Sigma_i (h_{i,0} - h_{i+j,0})^2 . \qquad (5.17)$$

6. Conclusion.

Our complaint about the nonsmooth character of Dirichlet distributed vectors is underscored in the continuous dimensional case. For the prior Dirichlet random process (Ferguson 1973), not only is the correlation structure unrealistic, but with probability one, the outcome distribution, or realization of the prior random process, is a discrete distribution, rather than a distribution having a density function, much less a smooth density. Realizations of the posterior process are again discrete distributions, and as in the prior process, the marginal distributions on partitions Π, $(\theta(S),\ S \in \Pi)$, are Dirichlet distributed, and hence nonsmooth. Our work with filtered-variate Dirichlet distributions can be viewed as a finite-dimensional version of Lo's (1984, 1987) prior process for Bayesian nonparametric inference. Lo used a Dirichlet process to mix over an infinite class of densities that are smooth to various extents. See also Antoniak (1974), Tiwari et al (1987), Escobar (1988), and the reply to discussants by Diaconis and Freedman (1986).

In our class of inference problems, the filter $\boldsymbol{\theta} = \mathbf{G}\boldsymbol{\alpha}$ is just used as a device in the construction of a tractable hierarchical prior distribution expressing opinion for a locally smooth $\boldsymbol{\theta}$. In another more general area of inference, $\mathbf{G}\boldsymbol{\alpha} = \Sigma \alpha_j \mathbf{g}_j$ might represent a

"mixed-distribution" sampling model with alternate conditional sampling distributions $\mathbf{g}_j$ and unknown mixing proportions parameter $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)^T$. At each trial in the sample, there would be an independent random selection of a distribution to draw from. For the case of known $\mathbf{G}$, interest would focus on the estimation of what would be a sampling parameter, $\boldsymbol{\alpha}$. Titterington, Smith, and Makov (1985, p 107) have proposed Dirichlet prior distributions for $\boldsymbol{\alpha}$ in finite mixture distribution problems.

The likelihood function of $\boldsymbol{\alpha}$ entering into our posterior density (3.1) is a product of weighted averages. In the special case where the prior support of $\boldsymbol{\alpha}$ is restricted to the set of vertices of the probability simplex ($c \to 1$, $\tilde{P}[\boldsymbol{\alpha} = \delta(j)] = u_j$, $j = 1, \ldots, m$), a mixture-distribution sampling model would randomly select an alternate model $\mathbf{g}_j$, once only, and maintain the same $\mathbf{g}_j$ for every trial in the sample. The single randomly selected model would be unknown, so one would have, in effect, a model-comparison inference problem. In this case, the likelihood would degenerate into a product of coordinates of $\mathbf{g}_j$, for alternative j, as usually seen in a Bayesian problem of comparison of hypotheses. The reader should avoid an unfortunate tendency to confuse this extreme model with our adaptation of the fuller mixture distribution as a sampling distribution in the particular class of smoothing inference problems.

More extensive experience is needed with the filtered-variate Dirichlet and other filtered-variate MSV prior distributions and their assessment. The mechanics of their prior assessment, updating to sample data, and posterior inference are easy enough, and their potential advantages are strong enough, we hope, to tempt the reader to experiment with their use.

We end by acknowledging that not all inference problems where the unknown category probabilities are known to be smooth call for smoothed estimates, that is, estimates that share sample information between neighboring categories. One of us (JMD) thinks he recalls an intriguing statement by John Tukey during informal discussion in 1971 to the effect that the nonsmooth appearance of a reported histogram can help warn against undue trust in a small sample.

7. References

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Annals of Statistics*, Vol. 2, 1152-1174.

Bar-Lev, Shaul K., and Enis, Peter (1986). Reproducibility and natural exponential families with power variance functions. *Annals of Statistics*. Vol. 14 No. 4, 1507-1522.

Bloch, Daniel A., and Watson, Geoffrey S. (1967). A Bayesian study of the multinomial distribution. *Annals of Mathematical Statistics*. Vol. 38, 1423-1434.

Cifarelli, Donato Michele, and Regazzini, Eugenio (1988). Distribution functions of means of a Dirichlet process. Typescript.

Diaconis, Persi, and Freedman, David (1986). On the consistency of Bayes estimates (with Discussion). *Ann. Statist.* Vol 14, 1-67.

Dickey, James M. (1968a). Estimation of disease probabilities conditioned on symptom variables. *Mathematical Biosciences,* Vol 3, 249-265.

Dickey, James M. (1968b). Smoothed estimates for multinomial cell probabilities. *Ann. Math. Statist.*, Vol 39, 561-566.

Dickey, James M. (1969). Smoothing by cheating. *Ann. Mathl. Statist.*, Vol 40, No 4, 1477-82.

Dickey, James M. (1973). Scientific reporting and personal probabilities: Student's hypothesis. *J. Roy. Statist. Soc., Ser. B,* Vol 35, 285-305.

Dickey, James M. (1983). Multiple hypergeometric functions: Probabilistic interpretations and statistical uses. *J. Amer. Statistical Assoc.* Vol. 78, No. 383 (Sept. 1983), 628-637.

Dickey, James M., Jiang, Jyh-Ming, and Kadane, Joseph B. (1987). Bayesian methods for censored categorical data. *J. Amer. Statist. Assoc.* Vol. 82, No. 399, 773-781.

Escobar, Michael David (1988). Estimating the means of several normal populations by nonparametric estimation of the distribution of the means. PhD dissertation, Dept. of Statistics, Yale University, New Haven, Conn.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, Vol 1, 209-230.

Good, I. J. (1950). *Probability and the Weighing of Evidence.* Hafner, New York.

Good, I. J. (1965). *The Estimation of Probabilities.* MIT Press, Cambridge, Mass.

Good, I. J., and Gaskins, R. A. (1971). Nonparametric roughness penalties for probability densities. *Biometrika* Vol 58, 255-277.

Good, I. J., and Gaskins, R. A. (1980). Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data. *J. Amer. Statistical Assoc.* Vol. 75, 42-56.

Hoffman, Mark S. (1987). *The World Almanac and Book of Facts 1987.* Pharos Book, New York, p 777.

Jaynes, Edwin T. (1979). Concentration of distributions at maximum entropy. In *E. T. Jaynes: Papers on Probability, Statistics, and Statistical Physics*, ed. by R. D. Rosenkrantz. Boston: Reidel 1983.

Jiang, Jyh-Ming (1984). Distributional properties of linear forms in a Dirichlet vector and applications. PhD dissertation, Res. Rept. 7/84-#14, Dept. of Mathematics and Statistics, State Univ. of N. Y. at Albany.

Jiang, Jyh-Ming (1988). Starlike functions and linear functions of a Dirichlet distributed vector. *SIAM J. Math. Anal.*, Vol 19, No 2 (March 1988), 390-397.

Lenk, Peter J. (1988). The logistic normal distribution for Bayesian nonparametric predictive densities. *J. Amer. Statist. Assoc.* Vol 83.

Leonard, T. (1973). A Bayesian method for histograms. *Biometrika.* Vol 60, 297-308.

Leonard, T. (1978). Density estimation, stochastic processes, and prior information (with Discussion). *J. Roy. Statist. Soc., Ser. B,* Vol 40, 113-146.

Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates: I density estimates. *Ann. Statist.* Vol 12, 351-357.

Lo, A. Y. (1987). Bayes methods for mixture models. Research Report No. 86-3. Dept. of Statistics, State Univ. of N. Y. at Buffalo, Amherst, N.Y.

Pagels, Heinz R. (1982). *The Cosmic Code: Quantum Physics and the Language of Nature.* New York: Bantam Books.

Tierney, L., and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *J. Amer. Statist. Assoc.*, Vol 81, 82-86.

Titterington, D. Michael (1986). A general view of statistical smoothing. *A.S.A. Proceedings of Business and Economic Statistics Section.* 79-84.

Titterington, D. M.; Smith, A. F. M.; and U. E. Makov (1985). *Statistical Analysis of Finite Mixture Distributions.* New York: Wiley.

Tiwari, Ram C.; Chib, Siddhartha; and Jammalamadaka, S. Rao (1987). Nonparametric Bayes prediction density estimation by random mixtures of multivariate distributions. Paper presented in 34th Meeting of the NBER-NSF Seminar for Bayesian Inference in Econometrics, Duke Univ., April 24-25, 1987.

U.S. Bureau of the Census (1987). *Statistical Abstract of the United States 1988* (108th ed) Wash. D.C., p 44.

Vardi, Y.; Shepp, L. A.; and Kaufman, L. (1985). A statistical model for positron emission tomography (with Discussion). *J. Amer. Statist. Assoc.* Vol 80, 8-37.

Whittle, P. (1958). On the smoothing of probability density functions. *J. Roy. Statist. Soc., Ser. B,* Vol 20, 334-343.